Statistics for Clinicians

Introduction

ETER C. O'BRIEN, Ph.D.
epartment of Medical Statistics
nd Epidemiology

MARC A. SHAMPO, Ph.D.
ection of Publications,
ivision of Education

Because the field of statistics has become so important in medicine, it is well worth a physician's time to become acquainted with the language of statistics, the elementary concepts, and a few of the more commonly used procedures. For this purpose, we are presenting these essentials in a series of 12 short, nontechnical papers. Obviously, what can be accomplished in the brief space allotted each of them is limited. However, the reader can expect to gain an understanding of what statistics is, an ability to understand the statistical aspects of much of the medical literature, a feel for when it will be necessary to consult with a statistician, and—for those occasions—an ability to communicate effectively with him.

Unfortunately, such an elementary acquaintance as we offer may lead a reader to overestimate his statistical capabilities and fail to consult a statistician in undertaking a research effort. We do not believe that this series of papers, or any review of statistics at the introductory level, will enable anyone to proceed without professional assistance in medical research requiring statistical expertise.

ORGANIZATION AND CONTENTS

Medical research studies may be classified into two broad categories. *Descriptive studies* are intended to describe the characteristics of only the study group, using observations obtained from every member of the group. *Inferential studies,* on the other hand, are designed to enable the investigator to use observations from selected individuals (a sample) to make conclusions about the larger group (population) from which they were drawn.

Our first three papers deal with descriptive studies, focusing on summary statistics (such as the mean and median) and graphic techniques (such as histograms and scatter diagrams). In paper 4, we describe how one may estimate characteristics of the population from characteristics of a small number of its members randomly selected. These principles are then applied in papers 5 through 8 to the problem of testing hypotheses about the population by use of some of the more common testing procedures.

Papers 9 through 12 discuss other common topics in medical research. Included are some problems that arise in analyzing survival data (where one must be careful to account for the fact that not all persons in the study were observed until death and some may have been followed up longer than others), determining normal values, evaluating a new medical procedure, and applying sequential statistical methods (which enable the investigator to test hypotheses while the study is in progress, with a view toward terminating the study early).

As we have indicated, our purpose is to offer an acquaintance with these topics for only a small investment of the reader's time. Throughout, the

discussion will be kept at an elementary level, omitting all mathematical derivations and, as much as possible, mathematical formulas. It is our hope that, upon completion of this series, the reader will be encouraged to go on to a further study of statistics. Many excellent elementary textbooks are available.

ACKNOWLEDGMENT

Although we assume all responsibility for the content of this series of papers, much of the organization framework is patterned after a course taught at the Ma Clinic for many years by Dr. Lila Elveback. We grateft acknowledge her contribution, including many hel conversations relating to these papers. Also we wish thank Dr. Guy Whitehead, who contributed much to writing and editing of the series.

1. Descriptive Statistics

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms and symbols introduced:
Interquartile range (P_{25} through P_{75})
Mean ($\bar{x}$)
Median (P_{50})
Outliers
Range
Skewness
Standard deviation (s)

Statistics is the mathematical technique or process of gathering, describing, organizing, analyzing, and interpreting numerical data. The study of statistics can be divided into two parts: descriptive and inferential. *Descriptive* statistics involves the numerical description of a particular group, whereas *inferential* statistics involves the process of taking a sample and making inferences about the population from which the sample was taken. In this first paper we will consider some elementary descriptive statistics.

TYPICAL VALUES

The two most important statistics for measuring typical values (that is, the location of the center of a set of data) are the mean and the median.

Mean.—The mean is computed by summing the individual data points, then dividing this sum by the number of observations (n) in the data set. We illustrate with the following hypothetical data:

$$-2,\ 0,\ 2,\ 4,\ 6\ (n = 5).$$

$$\text{The mean is } \frac{-2 + 0 + 2 + 4 + 6}{5} = \frac{10}{5} = 2.$$

Median.—If n is odd, the median is defined as the middle value: half the other observations are equal to it or smaller and half are equal to it or larger. For the data set $(-2, 0, 2, 4, 6)$, the median is 2. If n is even, one takes the midpoint between the two inner values: the median of (1, 5, 6, 7) is 5.5; and the median of (4, 10, 18, 36) is 14.

VARIABILITY

Regardless of which method (mean or median) has been used to locate the center of the data, the question of variability arises. Specifically, one is interested in the range of values that occur most commonly and how closely individual values tend to cluster around the center.

A useful method is to determine the 25th percentile (P_{25}) and the 75th percentile (P_{75}). Of all the values under consideration, 25% lie below P_{25} and 75% lie below P_{75}. The *interquartile range* (also called the semiquar-

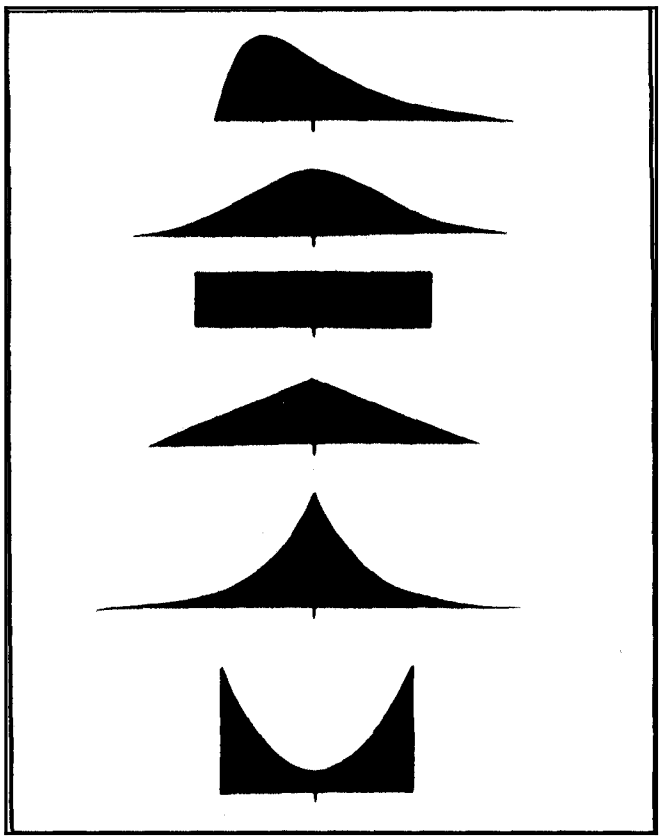

Fig. 1-1. Six data sets with same mean ($x = 4$) and same standard deviation ($s = 2.83$). (From Elveback LR: A discussion of some estimation problems encountered in establishing normal values. *In* Clinically Oriented Documentation of Laboratory Data. Edited by ER Gabrieli. New York, Academic Press, 1972, pp 117-137. By permission.)

tile range) extends from the value at P_{25} to the value at P_{75}, and this range includes 50% of the data points. In some instances, an investigator may find other percentiles more appropriate.

In very small data sets, an informative statement regarding variability is given by the *range*—the smallest value and the largest. However, a disadvantage of the range is that it depends heavily on the size of n: as more observations are included (as n becomes larger), the range usually gets larger (though it may remain unchanged). The range also may be greatly influenced by outliers, as will be illustrated below.

Another statistic that is commonly used to describe the variability in a set of data is the *standard deviation*. This usage of the standard deviation appears to derive largely from the mistaken belief that 95% of the observations can be expected to lie within two standard deviations from the mean. The falsity of this proposition is easily demonstrated, for it is true only under special, infrequently occurring conditions. Thus the appropriateness of the standard deviation for descriptive purposes is somew limited. However, it is useful in other contexts (relatin the sample mean) which will be discussed in later pap The computations required for calculating the stand deviation are illustrated below.

Step 1. Square the deviation of each individual va from the mean.

Step 2. Sum the squared deviations.

Step 3. Divide the sum by $n - 1$. The result is called *variance* (s^2).

Step 4. Obtain the standard deviation (s) by taking square root of the variance ($\sqrt{s^2}$).

Example

Step 1.

Original data	Deviation from mean of +2	Deviation squared
−2	4	16
0	2	4
+2	0	0
+4	2	4
+6	4	16

Step 2. Sum of squared deviations = 40

Step 3. ($n = 5$)

$$s^2 = \frac{\text{sum of squared deviations}}{n - 1} = \frac{40}{4} = 10$$

Step 4. $s = \sqrt{10} = 3.16$.

OUTLIERS AND SKEWNESS

Although the mean and standard deviation are the m commonly used statistics for describing typical val and variability exhibited by a set of data, they are appropriate when outliers or skewness is present. example, seven measurements of serum glutamic alacetic transaminase in the same subject produced values 8, 9, 9, 9, 10, 10, 20 units/ml. The value of clearly dissimilar to the other six observations, is ter an "outlier." When it is included, the mean is 1 which is larger than six of the seven data points. standard deviation, 4.2, is more than twice the rang the remaining six points when the outlier is omitt Clearly, in this instance, the mean and standard devia do not provide an accurate description of the set of d In this case, the data would be described more accura by a statement that the median value is 9, six values ra from 8 to 10, and one value is 20.

As an example of skewness, consider seven meas ments of serum triglycerides: 90, 93, 97, 103, 111, 1 153 mg/dl. The mean and standard deviation are 11 and 22.4, respectively. Note that the span from smal

value to median is only 13 units, 90 to 103, while the span from median to largest value is 50 units, 103 to 153. When the values are arranged in order of increasing size and those greater than the median are more spread out than those less than the median, we say the data are skewed to the right. This is a common occurrence, particularly with data that cannot be negative, such as the usual laboratory measurements. Less frequently, one encounters data that are skewed to the left.) Again the mean and standard deviation fail to represent accurately the typical values and dispersion. The median (103) and range (90 to 153) would convey this information better.

Generally, when data are highly skewed or when outliers are present, the center is more meaningfully measured by the median. Variability usually is best described by quoting appropriate percentiles or the range (or both), and this is especially appropriate when outliers or skewness is present. Ultimately, of course, the summary descriptive statistics discussed above remain a summary. Considerably more information may be conveyed by graphic displays.

Limitations of the mean and the standard deviation are illustrated by Figure 1-1,[1] which shows the manner in which individual values of six hypothetical data sets are distributed about the mean. (For example, notice that in the distribution at the top of Figure 1-1 most of the values are less than the mean, with the data skewed to the right.) Although the data sets depicted are very different, all six have the same mean ($\bar{x} = 4$) and same standard deviation ($s = 2.83$). A further discussion of graphic displays, which are especially useful in describing large data sets, will be the subject of our next paper.

REFERENCE

1. Elveback LR: A discussion of some estimation problems encountered in establishing normal values. *In* Clinically Oriented Documentation of Laboratory Data. Edited by ER Gabrieli. New York, Academic Press, 1972, pp 117-137

2. Graphic Displays—Histograms, Frequency Polygons, and Cumulative Distribution Polygons

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms and symbols introduced
Histogram
Class interval
Frequency polygon
Cumulative distribution polygon

This is the second in a series of papers dealing with the use of statistics in medicine. The previous paper discussed the use of *summary* statistics that describe a set of data by indicating their center (mean or median) and their variation from it (the standard deviation, range, or interquartile range). Also considered were the limitations of these statistics in describing the distribution of a data set. An illustration showed six very different distributions that all had the same mean (center) and standard deviation (variability). In this paper, graphic displays demonstrating the distribution of large data sets involving continuous variables will be considered.

Histograms.—A very useful graph for this purpose is the histogram, in which frequency is represented by area. For example, Figure 2-1 shows the distribution of serum triglyceride values from 96 6-year-old boys. It can be seen that there are more values in the interval from 41 through 50 mg/dl than in any other interval, and that most of the values are less than 71 mg/dl (the area to the left of 71 mg/dl is most of the total area). To provide an understanding of histograms, we will work through the steps that produced Figure 2-1.

The first step is to list the observations in order of size, indicating the frequency with which each observation occurs (Table 2-1). One then forms *class intervals*, grouping the data according to intervals of interest or in such a way as to ensure that each interval contains at least some minimal number of observations. On occasion, one may wish to use unequal class intervals. For example, in describing the age distribution of a group of subjects in which mortality is of interest, the first year of life may be of special interest; if so, class intervals 0-1, 2-9, 10-19, 20-29, and so forth may be desirable. To illustrate the technique for this expedient in the example involving the triglyceride values, unequal class intervals (columns A and B of Table 2-1) have been chosen, which will cause the columns in the histogram to be of unequal width.

In column C of Table 2-1 are the frequencies, or the number of observations that fall within each interval (in the example, the numbers of subjects whose triglyceride values fall within each interval). If all of the intervals were of equal size, these frequencies would suffice to determine the relative heights of the bars to be plotted in the histogram. Since the widths

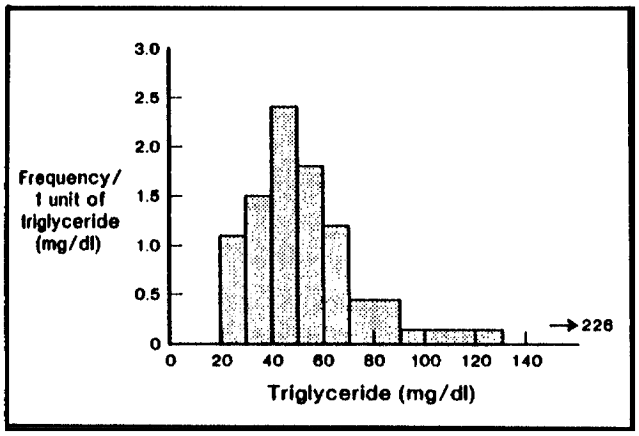

Fig. 2-1. Histogram of triglyceride values from 96 6-year-old boys corresponding to data in Table 2-1. Abscissa (x-axis) has unequal intervals corresponding to column *B* in Table 2-1. Ordinate (y-axis) has values corresponding to column *D* in Table 2-1.

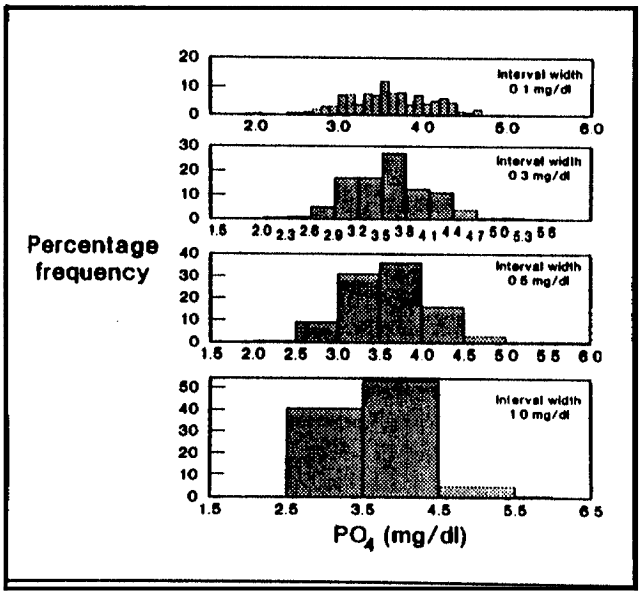

Fig. 2-2. Histograms of PO_4 levels in 329 females, plotted with interval widths of 0.1, 0.3, 0.5, and 1.0 mg/dl.

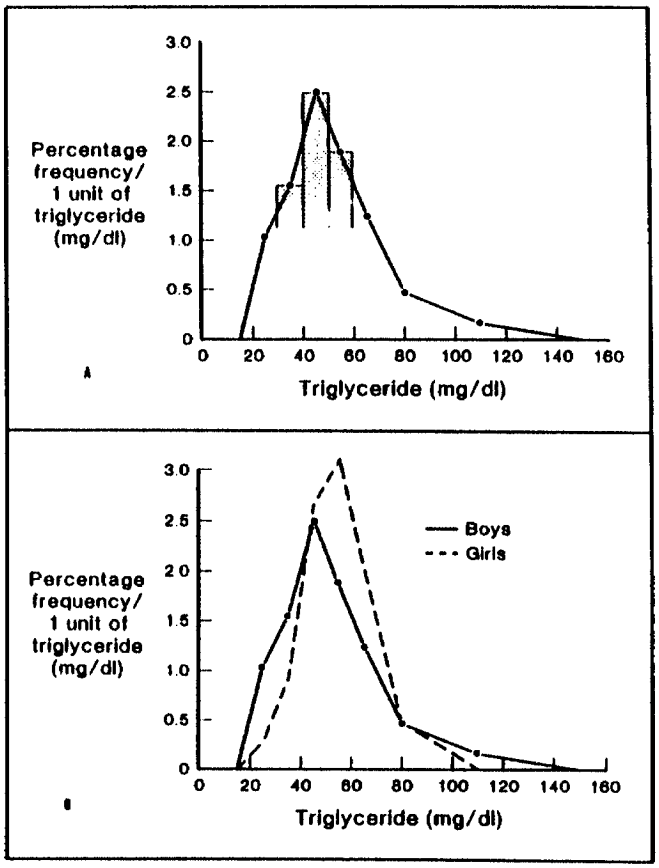

Fig. 2-3. Frequency polygons representing serum triglyceride values. Frequencies are expressed as percentage frequencies. *A*, Data from 96 boys (from Fig. 2-1 and Table 2-1). *B*, Data from 96 boys and 64 girls.

of the intervals are unequal and the frequency is to be represented by area (width × height), one must solve: frequency = width × height. Thus,

$$\text{height} = \frac{\text{frequency}}{\text{width}} = \text{frequency per unit of measurement}$$

(frequency per 1 mg/dl of triglyceride).

Table 2-1.—**Distribution of Serum Triglyceride Concentrations in 96 Boys 6 Years Old**

Triglycerides, mg/dl serum (A)	Width of interval (B)	Frequency (C)	Frequency ÷ width (D)	Cumulative % of subjects (E)
21-30	10	11	1.1	11.4
31-40	10	15	1.5	27.0
41-50	10	24	2.4	52.0
51-60	10	18	1.8	70.8
61-70	10	12	1.2	83.3
71-90	20	9	0.45	92.7
91-130	40	6	0.15	99.0
226	1	1	1.0	100.0

Except in the case of very large data sets, one must consider the problem of choosing interval widths, keeping in mind the twin objectives of accurate detail and reliable overall description of the distribution. These considerations are illustrated in Figure 2-2. Apparently, many of the peaks that are seen with use of 0.1 as the interval width are artifacts—notice that they disappear when an interval width of 0.3 is used. Conversely, with intervals of 1.0 virtually all detail is lost. However, no recommendation will be made for choosing between the two histograms in the middle (class intervals of 0.3 or 0.5) other than to point out that—as will often be the case—the informed judgment of the investigator will likely serve better than any rule of thumb.

Whatever class intervals are chosen, whether of equal or unequal width, the horizontal axis should be marked at regular intervals (like a ruler), as in Figure 2-1. The vertical axis should start at 0 and also be marked at regular intervals, and should not be broken.

Frequency Polygons.—Frequency polygons provide a useful method for comparing two data sets on the same graph. (If the sets are not of the same size, their distributions first are made proportional, usually by conversion to a percentage basis.) To draw a frequency polygon, one simply connects the midpoints of the tops of successive bars of the histogram (made with percentage frequencies), as shown in Figure 2-3 *A*. A frequency polygon comparison of triglyceride values from the 96 6-year-old boys in our previous example with the corresponding values from 64 6-year-old girls is shown in Figure 2-3 *B*.

Cumulative Distribution Polygons.—Another very useful method for displaying the distribution of a data set is provided by the cumulative distribution polygon (Fig. 2-4), which shows the percentage of observations less than any given value. Any desired percentile can be obtained from it as well. For example, Figure 2-4 indicates that among the set of 96 triglyceride observations in our familiar example, 80 mg/dl corresponds to the 88th percentile (88% of the observations were less than 80).

The graph is constructed by connecting consecutive points from the cumulative distribution (column E of Table 2-1) with straight-line segments. Cumulative frequency polygons can be plotted together, just as frequency polygons can; and this provides another way to compare sets of data.

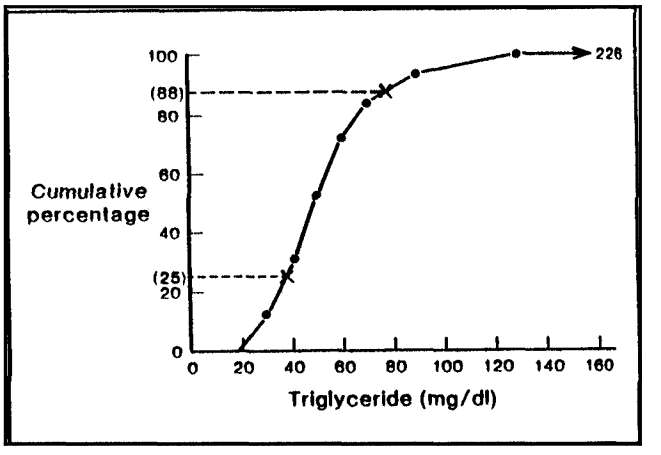

Fig. 2-4. Cumulative distribution polygon of triglyceride values from 96 boys (from column *E* in Table 2-1).

Statistics for Clinicians

3. Graphic Displays—Scatter Diagrams

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms and symbols introduced
Scatter diagram
Transformation

This is the third of a series of papers dealing with statistics used in medical research. The first discussed descriptive statistics that are useful in providing summary information about a set of data. This paper is a continuation of the second, which introduced some graphic displays for presenting data on a single continuous variable, such as serum triglyceride concentration. However, we now consider graphing the relationship between two continuous variables—for example, between age and serum IgE concentration.

The appropriate graph is a *scatter diagram* (Fig. 3-1). Each point in the scatter diagram is determined by two values. In our example, each patient will be represented by a single point whose location is determined by his age (on the horizontal scale) and his IgE value (on the vertical scale).

The first step in preparing a scatter diagram is to determine the range for each variable, so that the axes may be properly labeled. The graph should be approximately square, with no values plotted on the axes themselves. For a scatter diagram—unlike the graphs described in paper 2—it is not necessary to start either axis at 0.

A scatter diagram should be one of the first steps in data analysis. Data features that otherwise might go undetected may become obvious on the scatter diagram.

For example, in Figure 3-1 it is apparent that one subject (*arrow*) is considerably older than the others in the group. Also, there are more patients with IgE values below the mean value (286 ng/ml) than above it. With the use of the terminology introduced in paper 1, it can be said that the age of 70 years is an outlier and the data on IgE are skewed. As explained in that paper, these are important elements to consider in selecting appropriate descriptive statistics. For the present data, medians and ranges would be preferable to means and standard deviations.

This example illustrates a general rule that should always be kept in mind when displaying data graphically: *The purpose of a graph is to convey a quick visual impression.* Figure 3-1 accomplishes this by exposing the presence of outliers and skewness. However, it would be inappropriate to expect the reader to determine individual IgE measurements from such a graph, as that information could be obtained more conveniently from a table.

In paper 2 of this series we showed how two large data sets may be compared by use of frequency polygons. With smaller data sets, individual points may be plotted in a scatter diagram. For example, IgE values in

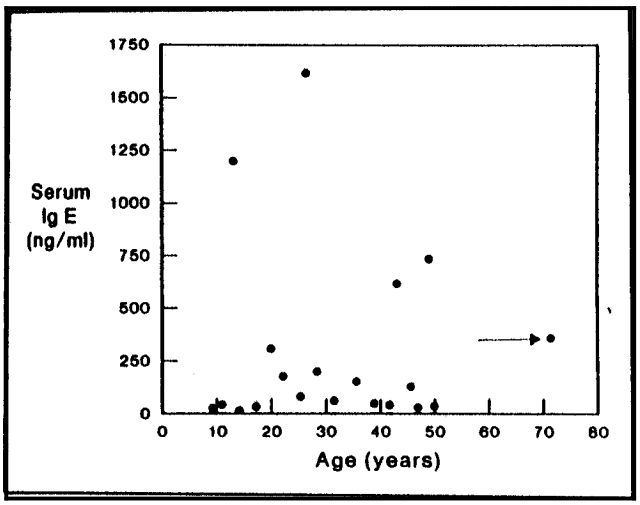

Fig. 3-1. Scatter diagram showing data (fictitious) relating IgE value with age.

males and females are compared in Figure 3-2 *A*.

When data are strongly skewed, as the data on Figure 3-2 *A* are, the display sometimes can be made more convenient by a suitable *transformation,* such as taking logarithms of the original measurements (Fig. 3-2 *B*). The same transformation may be accomplished simply by plotting the original values on semilog paper. Although the logarithmic transformation probably is the kind most commonly used, it is by no means the only one to be considered. Another transformation that is useful (especially when logarithms overcorrect, producing skewness in the opposite direction) is taking the square root of the variable.

If the transformation is successful in eliminating skewness, conceivably one could compute descriptive statistics (means and standard deviations) from the transformed data. Although this may be useful in some applications, it usually produces less satisfying results than would be obtained by choosing a more appropriate descriptive statistic that preserves the original unit of measurement. Generally, transformations are more useful in inferential than in descriptive statistics. The distinction between descriptive and inferential statistics will be the subject of the next paper.

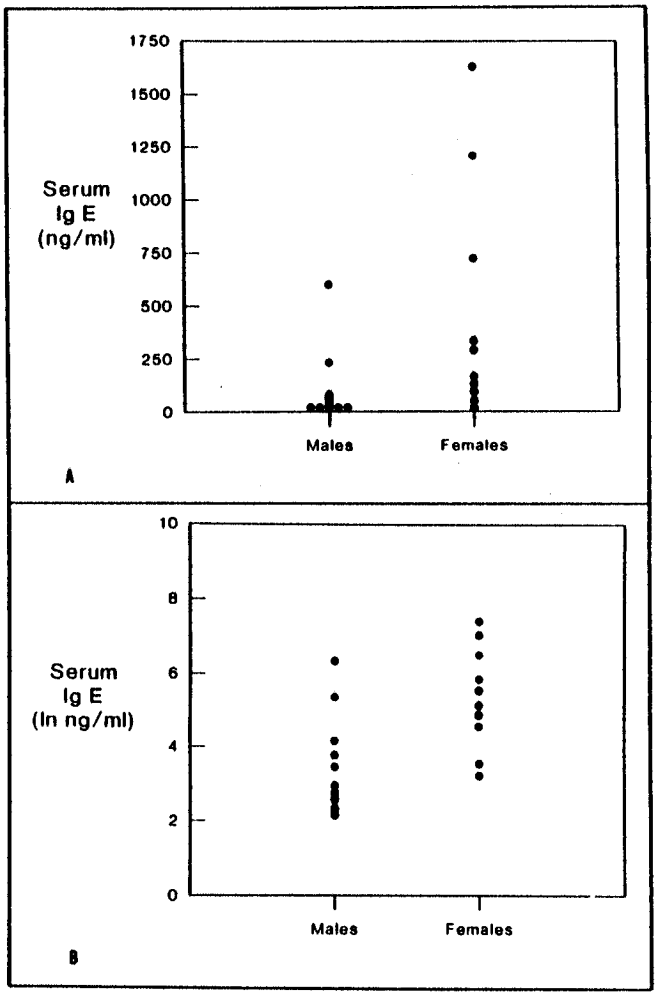

Fig. 3-2. IgE values by sex. *A*, Original measurements. *B*, Logarithms of measurements.

4. Estimation From Samples

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms and symbols introduced
- Population
- Random variable
- Parameter (μ, σ)
- Random error
- Standard error
- Confidence interval (confidence limits)

In the three preceding papers, we discussed statistical techniques for describing a set of data—descriptive statistics. Here we shall begin to consider *inferential statistics:* how to deal with problems wherein it is not practical to obtain and manipulate observations on every member of the population of interest. Our approach is to study a sample from the *population.* (Indeed, it is a convention of inferential statistics that "population" means a group—not necessarily of persons—which is studied by sampling.) To the extent that the sample group is representative of the population from which it is taken, inferences properly drawn from the sample will apply to the population.

DESCRIPTION OF POPULATION CHARACTERISTICS

Statisticians often refer to a population characteristic as a *variable;* for example, height, weight, and serum cholesterol would all be considered variables. The distribution of the values of a variable in the population can be represented by a sample histogram constructed with measurements from a sample group. Similarly, the sample mean and standard deviation ($\bar{x}$ and s) may be used to estimate the population mean and standard deviation (μ and σ). Statisticians refer to statistics such as $\bar{x}$ and s as *random variables,* since they vary randomly in repeated samples from the same population. The corresponding mean and standard deviation of the population—which are constants—are referred to as *parameters.* In distinguishing population parameters from their sample estimates, Greek symbols are generally used for the former and Latin symbols for the latter.

VARIABILITY OF RANDOM SAMPLES

The ability of sample statistics to describe population characteristics depends very much on the representativeness of the sample. To get some idea of the variation in random sampling (called *random error*), consider the data from a Mayo Clinic study of serum urea concentrations in 5,5[illegible] subjects.

Suppose that, having been provided with the values, we want to know their mean and standard deviation but do not want to add up 5,5[illegible]

Table 4-1.—Distribution of Serum Urea Values (mg/dl) in a Sample* (n = 100) Drawn Randomly From a Population (N = 5,594)

Value	Frequency	Value	Frequency
16	1	36	2
18	1	37	3
19	1	38	1
20	5	39	2
22	2	40	5
23	3	41	3
24	6	42	5
25	4	44	1
26	2	45	2
27	2	46	1
28	2	50	1
29	6	52	2
30	6	66	1
31	4	68	1
32	9	82	1
33	3	88	1
34	2	95	1
35	6	103	1
		173	1
			n = 100

*Mean of this sample ($\bar{x}$) is 36.56 and standard deviation (s) is 20.27; for calculations of $\bar{x}$ and s, see paper 1 (January 1981 issue).

numbers and do all the necessary further calculations on that large a scale. The 5,594 observations can be considered a population in the statistical sense and a random sample can be selected from it. Such a sample amounting to 100 observations is presented in Table 4-1, and a mean ($\bar{x}$) of 36.56 and standard deviation (s) of 20.27 have been calculated from it. In fact, when the necessary but tedious calculations were performed by a computer, the population mean (μ) was 35.33 and the population standard deviation (σ) was 21.55.

Since the samples drawn from a population vary, so do the estimates derived from them. To illustrate, we have drawn nine additional samples, each of size 100, from the population described above. As Table 4-2 shows, the means associated with the resulting set of 10 samples varied from 32.31 to 38.93.

ACCURACY OF SAMPLE MEAN AS ESTIMATE OF POPULATION MEAN

In judging how accurately a sample mean estimates the population mean, one begins with the realization that large samples are more reliable representatives than small ones. The procedures to be described here are suitable for samples containing as few as 60 observations, provided that the population does not have outliers or severe skewness.

With samples of sufficient size, regardless of the underlying distribution in the population, 95% of all sample means are within two *standard errors* of the population mean. The standard error of the mean ($SE_{\bar{x}}$) equals the standard deviation of the sample divided by the square root of the number of observations in the sample:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

So in sample 1 (Table 4-1), where s = 20.27 and n = 100,

$$SE_{\bar{x}} = \frac{20.27}{\sqrt{100}} = \frac{20.27}{10} = 2.03$$

And since the sample mean lies within two standard errors of the population mean in 95 of 100 instances, one can calculate the 95% *confidence interval (CI)* having the limits:

$$95\%\ CI = \bar{x} \pm 2 \cdot SE$$

Considered strictly, the "2" in the equation above is an approximation of a quantity that varies with sample size. But with n = 60 it is 2.00, and with extremely large samples it is 1.96; so when sample size is large, 2 usually is satisfactory. Continuing the application to sample 1, whose mean is 36.56:

$$95\%\ CI = 36.56 \pm 2 \cdot 2.03$$
$$= 36.56 \pm 4.06$$

Thus we can be confident, but not absolutely sure, that the population mean lies somewhere between confidence limits 32.50 and 40.62.

The 95% confidence interval provides a valuable indication of how much has been learned about the population mean from the sample. To obtain a narrower confidence interval, a larger sample is necessary. In the example above, if a confidence interval with a width of just 4 units instead of 8.12 (40.62 − 32.50) is desired, the sample will have to be increased to approximately 400 observations.

INFLUENCE OF SMALL SAMPLE SIZE

Thus far we have been using methods suitable for a moderately large sample. When the sample contains fewer than 60 observations, the number 2, by which we

Table 4-2.—Means of 10 100-Observation Samples From Population of 5,594 Serum Urea Observations

Sample no.	Sample mean (mg/dl)
1	36.56
2	33.92
3	34.24
4	33.00
5	35.47
6	36.67
7	35.15
8	38.93
9	32.31
10	36.57

multiply the standard error, must be replaced by a larger number (obtained from special tables). This number, which increases as sample size decreases, is designated by the symbol t^*_{n-1}.

Thus, when sample size is less than 60, decreasing the sample size increases the width of the confidence interval in two ways: (1) the standard error of the mean is increased, as illustrated in the previous section, and (2) t^*_{n-1} itself, the multiplier of the standard error, is increased. To illustrate, suppose that the standard deviation of 20.27 derived from the 100-observation sample had been obtained from a sample of only 10 observations. Then:

$$SE_{\bar{x}} = \frac{20.27}{\sqrt{10}} = \frac{20.27}{3.16} = 6.41$$

But also, the 95% confidence interval must be calculated thus:

$$95\%\ CI = \bar{x} + t^*_{n-1} \cdot SE$$

For the present sample ($n = 10$), $t^*_{n-1} = 2.26$. With this, and with the same mean obtained from sample 1 (36.56),

$$95\%\ CI = 36.56 \pm 2.26 \cdot 6.41$$
$$= 36.56 \pm 14.49$$

providing 95% confidence limits of 22.07 and 51.05.

So—despite retention of the same sample mean and standard deviation—the change from a basis of 100 observations to only 10 has changed the standard error from 2.03 to 6.41 and the width of the 95% confidence interval from 8.12 (40.62 − 32.50) to 28.98 (51.05 − 22.07).

COMMENT

1. Note that the standard deviation is not very helpful in describing the variability of the sample in Table 4-1. Specifically, the mean minus the usual two standard deviations becomes negative, which no actual serum urea value could be. As mentioned in the first paper of this series, the standard deviation has its greatest usefulness in relating sample means to population means—which is done by converting it to the standard error.

2. It will become more apparent in subsequent papers that much of the information required by statisticians in order to make probability statements is available only in special tables. Because the goal of this series is merely to acquaint the reader with basic concepts, the mechanics of working with the tables will not be discussed. It is hoped that the reader will not attempt to analyze his or her data, or even design the experiment, without the assistance of a statistician.

3. In this paper, we have dealt with the mean of a simple measurement, the serum urea concentration; and of course it might as well have been body weight or days of hospitalization. But further, the same concept of estimating a population mean from a sample—and for determining the confidence limits of the estimate—can be applied to differences (such as case-by-case differences in blood pressure before and after treatment) and to proportions (such as proportion of patients benefiting from a drug). The concepts presented here have very wide use in medical statistics. This will be illustrated in future papers of this series.

Statistics for Clinicians

5. One Sample of Paired Observations (Paired *t* Test)

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms and symbols introduced:
Paired *t* test
Null hypothesis
P value
Statistical significance
Paired observations

Paper 4 showed how a sample drawn from a large population can be used to provide information about characteristics of the population (such as its mean value) and also how the accuracy of such estimates can be assessed. In this paper, those methods will be applied in a procedure called the "paired *t* test" to solve a medical problem.

Formulation of the Problem.—The problem is to evaluate the effectiveness of a drug in lowering diastolic blood pressure. The population of interest consists of all patients who will receive the drug if it is used clinically in the future. The problem may be stated in two questions: (1) Will the drug reduce blood pressure? (2) If so, by how much?

With use of μ_{B-A} to represent the mean difference between measurements before *(B)* and after *(A)* treatment, if the drug is administered to the entire population as defined, the questions may be stated statistically: (1) Is $\mu_{B-A} = 0$? (2) If not, how large is μ_{B-A}?

Of course, it is not possible to determine μ_{B-A} directly by measuring the before-after difference in the total population of future patients. However, the methods described in the preceding paper can provide inferences about this parameter.

Collection of Data.—First, it is necessary to obtain a random sample from the population. Suppose only a very small pilot study consisting of 10 patients ($n = 10$) is to be done. If it can be assumed that patients present themselves in random order, the sample can be obtained simply by taking the next 10 patients who need treatment. Because it is rarely possible to conduct truly random collection in medical practice (as is often done in population surveys, for example), the question of the representativeness of the sample is an important aspect of any inferential study; but it will not be pursued in the present paper.

Suppose the sample is obtained appropriately, blood pressure is measured, the drug is administered, and blood pressure is measured again. In Table 5-1, note that two measurements are made on each patient. It is because these two measurements are made on the same patient and thus are correlated, not independent, that the data are regarded as a single sample of pairs.

Table 5-1.—Measurements of Diastolic Blood Pressure (mm Hg) Before and After Administration of an Antihypertensive Drug

Patient	Before (B)	After (A)	Δ (B-A)
1	110	111	−1
2	127	131	−4
3	132	138	−6
4	124	116	+8
5	118	117	+1
6	131	132	−1
7	104	99	+5
8	110	101	+9
9	126	120	+6
10	121	116	+5
	$\bar{\Delta}$ = +2.2 mm Hg		
	s = 5.1 mm Hg		
	$SE_{\bar{\Delta}}$ = 1.6 mm Hg		

In the sample, there is a mean decrease ($\bar{\Delta}$) of 2.2 mm Hg; and this serves as an estimate of drug effect in the population. It implies that the drug may reduce blood pressure.

However, that result is based only on sample data, subject to random error (which means that other samples from the same population probably would give different results). So one wonders: if there is no real difference between B and A (the *null hypothesis*), how often would a difference as large as 2.2 mm Hg occur in repeated samples from the population?

Question 1: Is $\mu_{B-A} = 0$?—The procedure is to make a probability statement of the sort, "If a given assumption or hypothesis regarding the population (such as $\mu_{B-A} = 0$) is true, then the probability of obtaining this sample result is no more than _____ (a value to be calculated)." And if the probability turns out to be sufficiently small, that will provide a basis for rejecting the hypothesis. In other words, when the sample result (an observed fact) is nearly impossible in conjunction with the hypothesis, one may reject that hypothesis in favor of an alternative hypothesis that seems more consonant with the data (for example, μ_{B-A} is greater than 0). It is important to recognize that all probability statements are "If . . . then . . ." statements, expressing the probability that, under carefully stated circumstances, something will happen or be true.

In the present example (Table 5-1), the first step toward determining the quantity needed for completing the probability statement is to calculate the size of the mean difference relative to the standard error of the difference. If the pair-by-pair differences include no outliers or evidence of severe skewness, the following formula may be used:

$$t = \bar{\Delta}/SE_{\bar{\Delta}}$$

(Notice that the variation associated with $\bar{\Delta}$, which is $SE_{\bar{\Delta}}$, is based on the variation among the pair-by-pair differences.) Substituting from Table 5-1,

$$t = 2.2/1.6 = 1.375$$

And, using special tables or computing facilities, we find that, *if* $\mu_{B-A} = 0$, *then* the probability of obtaining a value for t greater than 1.375 is 0.101. This probability is often referred to as a P value; so here, $P = 0.101$. It means that the observed difference would occur by random variation (without an underlying real difference) in 10.1% of samples.

The interpretation of this probability must be clear and not careless. What can we say?

1. We cannot reject the hypothesis $\mu_{B-A} = 0$. Since the observed results would occur fairly often even if the drug had no real effect, that may be the case—no real effect.

2. Conversely, we cannot rule out the possibility that a real effect exists, since a real effect might have gone undetected because of the small sample size. We can say only that the evidence in favor of a real decrease is not statistically significant.

Question 2: How Large Is μ_{B-A}?—In this situation, it is of interest to ask, "What values of μ_{B-A} are consistent with the observed results of our study?" The methods described in paper 4 can provide a 95% confidence interval for μ_{B-A}:

$$95\%\ CI = \bar{\Delta} \pm t^*_{n-1} \cdot SE$$

The value of t^*_{n-1} is obtained from a standard statistical table; and for the present example ($n = 10$), $t^*_{n-1} = 2.26$. Thus,

$$95\%\ CI = 2.2 \pm 2.26 \cdot 1.6$$
$$= 2.2 \pm 3.6$$

So we may be confident that the interval from −1.4 to +5.8 contains the true value of μ_{B-A}. The confidence stems from the fact that intervals constructed by this method contain the true value in 95% of trials with different samples. Obviously, the result obtained in our small sample could have occurred with no real underlying difference or with a sizable positive real difference (blood pressure decreased) or even a negative real difference (blood pressure increased).

Was the Sample Large Enough?—In general, confidence intervals are very useful in assessing the adequacy of sample size. If an effect exists, the harder we look for it the more likely we are to find it. A wide confidence interval says that we have not tried very hard (have not examined a large enough sample); and in that circumstance, failure to produce a small P value should not be regarded as demonstration that no effect exists.

To illustrate this point further, suppose that in the previous example the same mean decrease ($\bar{\Delta} = 2.2$) and standard deviation ($s = 5.1$) had resulted from a sample of size $n = 100$. In this case, calculations similar to those described above reveal that $P = 0.001$, indicating that (if

there is no real difference) random variation would produce the observed effect only 1 time in 1,000. Similarly, the 95% confidence interval becomes 1.2 to 3.2—much narrower than with the original small sample and no longer including 0.

COMMENT

1. One might ask, "How small a *P* value is required to achieve *statistical significance?*" The answer to this question depends on the circumstances of the particular study and, in general, it is best not to think in terms of black and white—significant or not significant. However, for guidelines one may consider *P* values between 0.10 and 0.05 as suggestive of a difference, though not statistically significant. The term "statistically significant" is usually reserved for situations where *P* is less than 0.05; and often the evidence of a difference is not considered conclusive unless the *P* value is less than 0.01.

2. Although the evidence of a drug effect in the present example, with $n = 100$, would be described as statistically significant (not likely to occur in the absence of a drug effect), the more important question—is it clinically significant?—is still unanswered. Whereas the statistician can help in addressing this very important question by providing confidence limits, as in the example, the ultimate decision must come from the clinician.

3. To provide the *paired observations* for the paired *t* test, each item in one data set must have an intrinsic correspondence with one—and only one—item in the other set. "Before" and "after" measurements from the same person (as in our example) are a frequent source of paired data. Pairing of data from different persons may be appropriate if the persons have been carefully matched. For instance, in comparing the effects of two drugs, an investigator might exclude genetic variation by using twins—giving drug X to one and drug Y to the other. The resulting paired data would be analyzed as in our example. More commonly, there may be two or three factors with major influence on response to treatment, making it desirable to recruit subjects in pairs—the members of each pair being similar to each other with respect to the factors identified as most important. Then, after one member of the pair is treated and the other is not, an observed difference between them should reflect response to treatment.

4. The ways data can be analyzed are determined by the way the study was designed. In point of fact, choosing the appropriate study design so as to be able to answer the questions of interest—and do so efficiently—is far more important, and also more difficult, than choosing the appropriate method of analysis.

Throughout this paper, many design considerations have been omitted in the interest of keeping things simple. The reader who is familiar with clinical trials for evaluating drug effectiveness probably has asked the question, "Shouldn't the study be double-blind with controls?" These are important considerations indeed, and they will be addressed in the next paper.

Statistics for Clinicians

6. Comparing Two Samples (The Two-Sample *t* Test)

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms and symbols introduced
- Control group
- Placebo
- Two-sample *t* test
- Relative deviate
- Randomization
- Double-blind

Paper 5 described a hypothetical study evaluating the effectiveness of an experimental drug in reducing diastolic blood pressure. The study design consisted simply of obtaining measurements before and after the administration of the drug in a series of consecutive cases. Although this type of study is satisfactory for many research objectives, it is not adequate for full assessment of the effectiveness of a drug. Specifically, we want to know whether an observed reduction of blood pressure was due to a biologic drug effect or—in whole or in part—to a psychologic response of the patients receiving the medication.

Study Design and Data Collection.—In order to isolate this psychologic response, we can include in the study a *control group*. These patients are like the others, but they are to be given only a *placebo*—that is, a preparation that resembles the experimental drug in all outward respects but has no biologic capability affecting blood pressure. The result obtained in the experimental group will be compared with that obtained in the control group.

Notice that, although both samples have been obtained from the same population, they are two distinct, independent samples. Further, there is no one-to-one matching between the individual subjects in the two samples. Therefore a pair-matched analysis like the one used in paper 5 would be inappropriate. A more suitable method (among several) is the *two-sample t test*.

Suppose the average change of diastolic blood pressure (from before to after treatment) in the experimental group is a decrease of 3.2 mm Hg with standard deviation 5.1 ($\bar{\Delta}_E = 3.2, s_E = 5.1$), whereas the average change in the control group is a decrease of 0.5 with standard deviation 4.0 ($\bar{\Delta}_C = 0.5, s_C = 4.0$).

Is There a Difference?—Although the apparent effect of the drug is greater than that of the placebo (3.2 versus 0.5), we ask the familiar question: What is the probability of obtaining such an apparent difference of effectiveness between the drug and the placebo? Assuming that the data contain no outliers or severe skewness, and noting that the standard

deviations are similar, we compare the mean difference between groups to the variability present in both groups:

$$t = \frac{\bar{\Delta}_E - \bar{\Delta}_C}{s_{pooled}\sqrt{\frac{1}{n_E} + \frac{1}{n_C}}}$$

in which:

$\bar{\Delta}_E - \bar{\Delta}_C$ = difference between mean change in experimental group and mean change in control group

n_E = number of patients in experimental group (receiving drug)

n_C = number of patients in control group (receiving placebo)

s_{pooled} = a combination of the standard deviations of the two groups

(Note that the denominator $s_{pooled}\sqrt{1/n_E + 1/n_C}$ is analogous to the denominator in the equation for t in paper 5. It is the standard error of the difference in the numerator. In all of our examples using the t test, no matter how complicated the equation becomes—how many factors or symbols are included—we are still computing a *relative deviate*, dividing the numerator by its standard error.

If we suppose that there were 100 patients in each group, computation of s_{pooled} gives 4.58; and appropriate substitutions yield:

$$t = \frac{3.2 - 0.5}{4.58\sqrt{\frac{1}{100} + \frac{1}{100}}} = 4.17$$

From suitable tables or computing facilities we find that, if there were no difference between the effect of the experimental drug and that of the placebo, a value of t as large as 4.17 would be obtained from 1.8% of repeated experiments ($P = 0.018$). Thus the data are not consistent with the hypothesis of no drug effect at the $P = 0.018$ level. So we reject the null hypothesis: the observed result is so unlikely to occur without a real underlying difference that there almost certainly is such a difference.

How Much Difference?—As in paper 5, the next question is: How much drug effect do these data imply? In this example, the 95% confidence interval for the true mean difference (drug effectiveness) is given by:

$$95\%\ CI = (\bar{\Delta}_E - \bar{\Delta}_C) \pm t^* \cdot s_{pooled} \cdot \sqrt{\frac{1}{n_E} + \frac{1}{n_C}}$$

which, with appropriate substitutions, becomes:

$$95\%\ CI = (3.2 - 0.5) \pm 2 \cdot 4.58 \cdot \sqrt{0.02}$$
$$= 2.70 \pm 1.30$$
$$= 1.40 \text{ to } 4.00$$

Although we cannot be certain that this interval contains the true difference, the method used to obtain the confidence interval is one which leads to an interval that does contain the true difference in 95% of applications.

As in the previous paper, the most important question must now be addressed by the physician: Is the effectiveness of the drug significant clinically?

COMMENT

1. This example illustrates the importance of including a control group in a clinical trial. If only the 100 patients given the experimental drug had been considered, without a control group given the placebo, a paired t test would have demonstrated a statistically significant difference as it did in paper 5. But although the analysis would have been valid mathematically, it could not have given any indication as to whether the observed result was due to the biologic effect of the drug.

2. In practice, each patient in this study would have been assigned to drug or placebo treatment by a process of *randomization*. For example, the drug manufacturer—using a randomization chart prepared by the project statistician—might provide the physician with 200 numbered envelopes, each containing the substance to be given to a different patient. Neither the physician nor the patient would know the contents of any envelope (such studies are referred to as *double-blind*), but this information would be recorded and retained by both the statistician and the manufacturer.

From the research standpoint, a randomized, double-blind study design is desirable because of the need to avoid biases in patient selection and assignment and in measurement of the response to treatments—thus to avoid compromising the basis for probability statements in the subsequent analysis. From the patient-care standpoint, the undesirable aspects of assigning treatment to patients randomly and double-blind are obvious. The justification of such management often rests on assumptions that the experimental treatment is not known to be superior (or inferior) to the placebo and that this information can be obtained only from a properly designed study.

3. In addition to the design considerations mentioned thus far, numerous others would need to be considered in developing a research protocol. In our example, we might wish to reduce the large variation inherent in measurement of blood pressure by taking multiple readings. It would also be desirable to standardize the conditions under which measurements are made, to ensure their comparability. Restricting the age range or the range of initial levels of diastolic blood pressure so as to obtain more homogeneous samples might be advantageous. In short, many factors would need to be considered, each requiring the close cooperation of clinician and statistician before the data are collected.

Statistics for Clinicians

7. Regression

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms and symbols introduced
- Least-squares line
- Linear regression equation
- Intercept (*a*)
- Slope (*b*)
- $s_{y \cdot x}$
- Standard error of *b* (SE_b)
- Correlation coefficient (*r*)

Recent papers have dealt with a hypothetical problem of evaluating the effectiveness of a drug for reducing diastolic blood pressure. When a drug is found to be efficacious, we want to know what influence other factors may have upon its effect. Specifically, in this paper, we shall measure the influence of initial diastolic pressure on the reduction achieved by the drug.

Suppose the reductions of diastolic pressure in 10 study participants were as listed in Table 7-1. The first step in analysis is to exhibit the data graphically, relating the changes to initial values by use of a scatter diagram (Fig. 7-1).

In order to quantify and summarize the association shown by the scatter diagram, we draw a straight line through the group of points, as illustrated in Figure 7-2. How well the line fits the data is measured by the sum of the squared vertical distances of the individual points from the line. Thus the best-fitting line is the one for which this sum of squares is least, and it is called the *least-squares line*. (There is a formula for the calculation.)

In general terms, the least-squares line may be described by the equation:

$$y = a + bx$$

This is the *linear regression equation*, in which:

a = *intercept*, the point on the y axis where the regression line will cross it, if extended that far (the value of *y* when $x = 0$).

b = *slope*, the amount of change in *y* per unit of increase of *x*.

For the line shown in Figure 7-2, $a = -23.53$ and $b = 0.4671$; so the specific equation is:

$$y = -23.53 + 0.4671x$$

We are also interested in measuring how closely the points cluster about the regression line. The appropriate measure, denoted by $s_{y \cdot x}$, is defined in terms of the sum of the squared vertical distances from the regression line (as shown in Figure 7-2). Specifically, $s_{y \cdot x}$ is defined as the square root of this sum after dividing by *n-2*. Rather than assign a special name to this statistic, statisticians usually write the symbol itself and pronounce it "s-y-dot-x." In this example, $s_{y \cdot x} = 2.257$.

In most applications, the feature of greatest interest is the slope, which here represents the amount of post-treatment change in pressure that

corresponds to a unit increase in initial pressure. Although we cannot determine definitely from a sample the magnitude of the unknown true slope (usually represented by β) in the population, we can estimate it from the sample, test hypotheses about it, and establish confidence limits as we did in the previous papers concerning population means.

Specifically, in our example, the true slope β is estimated by the sample slope b (0.4671). To test the hypothesis that $\beta = 0$, we begin by comparing b to its standard error (SE_b)—which depends on how closely the sample points cluster about the fitted line. In the example, $SE_b = 0.0751$, so the test statistic is:

$$t = b/SE_b$$
$$= 0.4671/0.0751$$
$$= 6.220$$

From suitable tables or computing equipment, we find that, if $\beta = 0$, then the probability of obtaining a value of t as large as 6.220 is less than 0.001 ($P<0.001$). So from our hypothetical data, we reject the hypothesis $\beta = 0$ and conclude that the drug response does depend on initial blood pressure.

The 95% confidence interval is given by:

$$95\%\ CI = b \pm t^*_{n-2} \cdot SE_b$$

where t^*_{n-2} is a number obtained from special tables. In this case $t^*_{n-2} = 2.306$. Upon substitution,

$$95\%\ CI = 0.4671 \pm 2.306 \cdot 0.0751$$
$$= 0.4671 \pm 0.1732$$

So the interval is from 0.2939 to 0.6403—well above zero, thus confirming our rejection of the hypothesis that the initial level and the amount of change were unrelated ($\beta = 0$). The data indicate that the initial level and the amount of change are related.

COMMENT

1. We have shown how the association between two variables may be quantified by fitting a straight line to the data. In doing so, we have considered only the simplest of situations. In practice, other factors may require attention: for example, how to modify the analysis if the scatter diagram reveals outliers or skewness or associations that are nonlinear, or how to evaluate additional variables (such as sex or obesity).

Two other considerations regarding the regression line should be remembered. First, in graphing the regression line the steepness of the line depends on how the axes are scaled (whether large or small units are used). Second, extension of the regression line beyond the plotted data may give rise to absurd implications.

2. You may have noticed that we have not mentioned a rather popular statistic called the *correlation coefficient*, usually denoted by r. Its popularity derives in part from the fact that the correlation coefficient does not depend on the units of measurement (for example, pounds or kilograms) as the slope of the regression line does. The correlation coefficient is a somewhat com-

Table 7-1.—**Data for Hypothetical Example**

Subject	Pressure before drug (B)*	Pressure after drug (A)	Difference (B - A)
1	104	80	24
2	110	81	29
3	110	81	29
4	111	85	26
5	118	83	35
6	124	93	31
7	126	91	35
8	127	89	38
9	131	95	36
10	132	93	39

*Initial diastolic blood pressure (mm Hg).

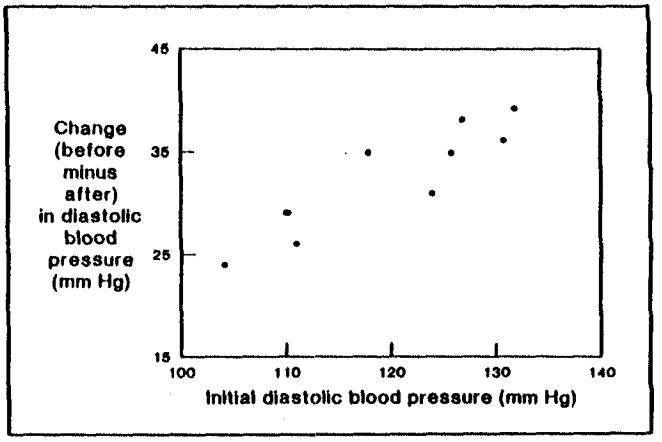

Fig. 7-1. Scatter diagram of initial diastolic blood pressure and change in diastolic blood pressure, from Table 7-1.

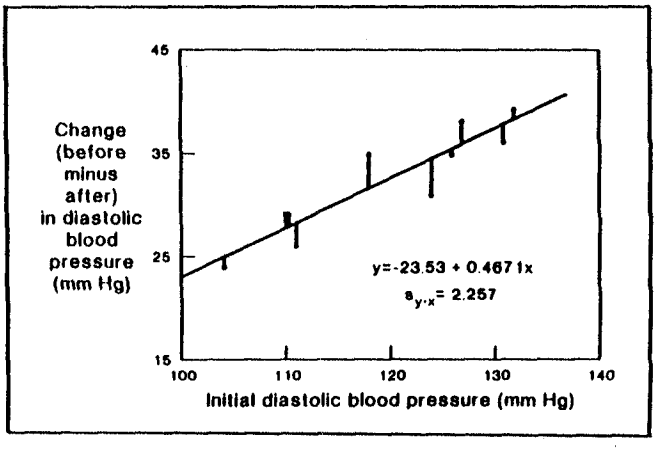

Fig. 7-2. Scatter diagram (Fig. 7-1) with regression line and lines from data points to regression line for least-squares determination.

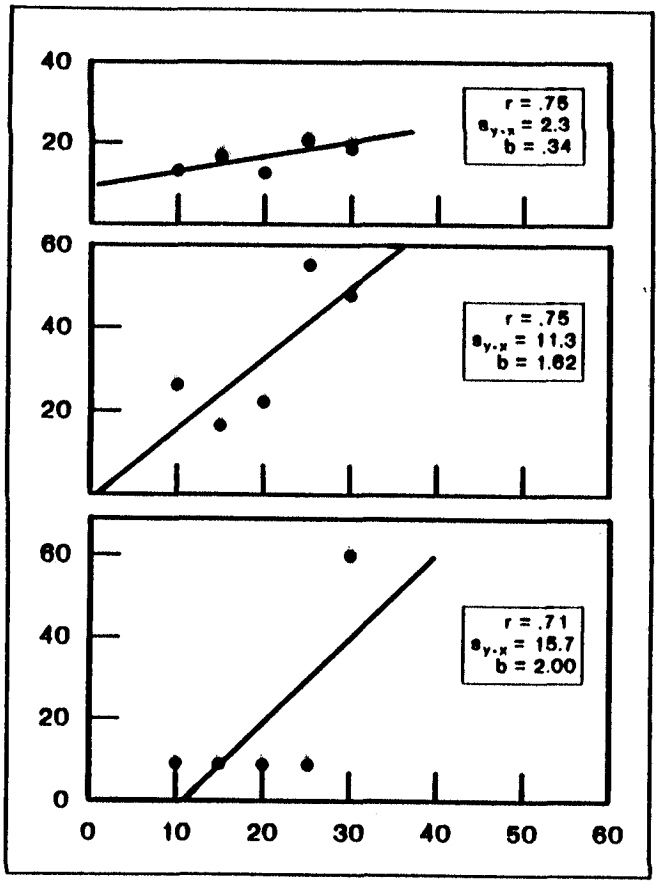

Fig. 7-3. Examples of similar correlation coefficients resulting from different conditions. *Top Panel,* Smallness of scatter about regression line. *Middle Panel,* Steepness of slope. *Bottom Panel,* Presence of outlier.

plicated function of the test statistic b/SE_b and the sample size. The sign of b (that is, whether the regression line runs upward or downward) determines whether r will be positive or negative; and when $b = 0$, $r = 0$. The calculation of r can be used to test the hypothesis that y is *not* related to x.

However, if y *is* related to x, r serves poorly in describing how, because it is ambiguous. Although a large value of r (within its mathematical limits of +1 and −1) suggests that the correlation is strong, faith in this simple implication may be misplaced. The value of r can be increased by increase of b and also by decrease of $s_{y\cdot x}$ (since SE_b is directly proportional to $s_{y\cdot x}$). These components are quite different: b (the slope of the regression line) indicates how large an associated change is; and $s_{y\cdot x}$ (the closeness of the data points to the regression line) indicates how consistently the change occurs. But r, as a single value, gives no indication of the relative influence of the two components in determining its value.

Notice in Figure 7-3 that, although the correlation coefficient is virtually the same in each instance, the associations between y and x are much different. The high correlation coefficients are due, successively, to smallness of the scatter about the line, to steepness of the slope, and to presence of an outlier. These examples also illustrate the importance of looking at a scatter diagram whenever one does a regression analysis.

To describe the association between two variables in terms of summary statistics, it is best to use both b and $s_{y\cdot x}$.

Statistics for Clinicians

8. Comparing Two Proportions: The Relative Deviate Test and Chi-Square Equivalent

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms introduced
- Dichotomous variable
- One-sided test
- Two-sided test
- Relative deviate

Paper 7 presented a method for comparing observations of two continuous variables. Such variables are called "continuous" because they can have a continuum of values; and the measurement of interest in paper 7 was blood pressure.

FORMULATION OF THE PROBLEM

In this paper, we consider how to compare *dichotomous variables,* which are observed as yes-no, alive-dead, normal-abnormal, and so on. For an example, let us compare the incidence (yes-no) of a side effect (headache) in association with each of two drugs: 15 of 50 cases with drug F and 8 of 50 cases with drug G.

Note that the dichotomous observations of each group can be summarized by a proportion, which will express the incidence within that group as a degree on a continuous scale of possibilities. Let π_F and π_G represent the proportions of the incidence of headache associated with drugs F and G, respectively, in the populaton—the true (but unknown) proportions. For an estimate of π_F, , we can use the sample proportion $p_F = 0.30$ (15/50); and for π_G, we can use the sample proportion $p_G = 0.16$ (8/50).

Employing these terms, we state the familiar two questions: (1) Is there a real difference between these groups—that is, does $\pi_F = \pi_G$? and (2) How large may the difference be?

QUESTION 1: IS THERE A DIFFERENCE?

If $\pi_F = \pi_G$ (if the proportions π_F and π_G are the same), we can write this unknown common proportion as π_0. To obtain a corresponding sample statistic (p_0) in accord with the null hypothesis that there is no underlying difference between the samples (that the apparent difference is only random variation), we pool the samples:

$$p_0 = \frac{15 + 8}{50 + 50} = 0.23$$

This resulting value of 0.23 is an estimate of the common proportion assumed (for test purposes) to satisfy the hypothesis in the question.

Again we compute the ratio (here we use the test statistic z) of the difference between the two data sets to the standard error of the difference (the variability within each data set as calculated with the sample statistic p_0). Still assuming that the null hypothesis is true (no difference between the samples), we use the common proportion p_0 in the denominator for this calculation.

$$z = \frac{p_F - p_G}{\sqrt{p_0(1-p_0)\left(\frac{1}{n_F} + \frac{1}{n_G}\right)}}$$

$$= \frac{0.30 - 0.16}{\sqrt{0.23(0.77)\left(\frac{1}{50} + \frac{1}{50}\right)}}$$

$$= 1.663$$

In this example, we will reject the null hypothesis ($\pi_F = \pi_G$) if either drug is found to cause fewer headaches than the other. (This differs from the interpretations in the two preceding papers. There we asked, a priori, "Is A superior to B?" Here we are asking, "Is either F or G superior to the other?") Hence we look for the probability of getting a value of z that is either 1.663 or higher (signifying more headaches with drug F) or −1.663 or lower (signifying more headaches with G). From appropriate tables, this probability $P = 0.096$; so we remain unsure that either drug excels the other in regard to incidence of headache.

QUESTION 2: HOW LARGE MAY THE DIFFERENCE BE?

An approximate 95% confidence interval for $\pi_F - \pi_G$ can be calculated with this formula:

$$95\%\ CI = p_F - p_G \pm 1.96 \cdot \sqrt{\frac{p_F(1-p_F)}{n_F} + \frac{p_G(1-p_G)}{n_G}}$$

Note that because the confidence interval will contain values of π_F and π_G that are unequal, we can no longer use p_0 in our estimate of the standard error.

$$95\%\,CI = 0.14 \pm 1.96 \cdot \sqrt{\frac{0.30 \cdot 0.70}{50} + \frac{0.16 \cdot 0.84}{50}}$$

$$= 0.14 \pm 0.163$$

Thus the 95% confidence limits are −0.023 and +0.303.

Even though the P value is greater than 0.05 (that is, 0.096) and the 95% confidence interval for $\pi_F - \pi_G$ contains 0, we still might conclude that the data provide suggestive evidence of a superiority for drug G. Our large confidence interval (reflecting the somewhat small sample size) indicates that drug G may offer a substantial advantage despite the lack of statistical significance.

It is a convention that P values are to be considered significant only if they are less than 0.05, and some investigators require P values less than 0.01 for convincing evidence against the null hypothesis. However, the distinction between significant and nonsignificant test results depends on circumstances in the individual study; and often an intermediate interpretation is appropriate, as it is this time. More generally, a P value should be interpreted as a measure of the strength of the evidence against the null hypothesis. Such strength can have many degrees, and it offers more meaning than "enough" and "not enough."

COMMENT

1. An additional lesson is concealed in this example. Suppose the investigators had not thought carefully about the problem of associated headaches until they saw that more occurred with F than with G. They might have formulated a hypothesis that G was superior in this regard and tested it looking only for a difference in one direction. The outcome would have been a statistically significant superiority for drug G ($P = 0.048$).

What is the probability of the investigators making a mistake when they take this approach to hypothesis testing? Let us suppose that there is no real difference between F and G. The probability of erroneously concluding that G is superior is 0.048. However, it is equally likely that the sample results would favor F by the same amount; and this also would give $P = 0.048$. Thus the probability for error is the probability of concluding G superior to F plus the probability of concluding F superior to G, which is 0.048 + 0.048 = 0.096.

In general, how do we determine whether to look for differences in just one direction (a *one-sided test*) or in both directions (a *two-sided test*)? The answer is to formulate the hypothesis clearly, and *before* the data are collected. The way the hypothesis is stated will determine how the test should be done. For example, when we ask the question "Is experimental drug A superior to placebo?" we clearly are looking for a difference in only one direction. If the experimental drug is found to perform either the same as or worse than placebo, the same negative conclusion will be reached. Since we are not interested in establishing that A is worse than placebo, a one-sided test is appropriate. (Notice that this was the situation in our previous examples, where all our tests were one-sided.)

Conversely, when comparing two drugs (as in the present example), we may ask: "Is either drug superior to

the other?" In this instance, we clearly are interested in establishing superiority in either direction, so a two-sided test is appropriate.

The decision as to whether a test should be one-sided or two-sided illustrates a very important principle in statistics: the study objectives and specific hypotheses to be tested should be formulated *before* the data are collected.

2. Various computational formulas are available for performing the test described in this paper. Since they all give the same *P* value, they are equivalent. The formula that is simplest computationally and is used most commonly is called the chi-square (χ^2) test. (The number actually computed is z^2.) Although we have presented the computations in terms of the *relative deviate* statistic in order to provide a better understanding of the test, in practice, tests for comparing two proportions are most commonly referred to as chi-square tests.

3. In our previous examples, the computed test statistic was usually denoted by the letter *t*. Although any letter could have been used, *t* ordinarily is chosen for those situations because it corresponds to the name of the statistical tables used in obtaining the related *P* values. For the tables used in the relative deviate test for comparing two proportions, the letter *z* is commonly used. When the test is based on the simple computational formulas (which yield the square of the relative deviate *z*), the test statistic is denoted by the symbol χ^2.

4. After presenting methods for describing a set of data in papers 1-3, we introduced the concept of inferential statistics in paper 4 and provided examples in papers 5 to 8. In the remainder of this series, we shall focus on some of the most common topics that arise in medical research: evaluating a new diagnostic procedure, sequential methods, survivorship studies, and normal values.

Statistics for Clinicians

9. Evaluating a New Diagnostic Procedure

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms and symbols introduced
Reliability
Accuracy
Coefficient of variation

When a new medical procedure has been developed, such as computer-assisted scanning, it is necessary to evaluate the contribution to patient care that will result from its use. In this situation, the subjective opinion of the physician responsible for patient care will be essential, and perhaps it will determine the ultimate decision as to the procedure's usefulness. However, it is desirable also to perform studies that will provide objective, quantitative data. Three aspects that should be considered are (1) the reliability of the procedure, (2) its accuracy, and (3) how it compares with conventional methods.

We shall use evaluation of an experimental computer-assisted scanner to illustrate how each of these concerns may be addressed. The statistical methods employed will differ slightly, according to whether the measurement of interest is dichotomous (such as presence or absence of a tumor) or continuous (such as tumor size). We shall consider the dichotomous type first.

DICHOTOMOUS DATA

Reliability.—The reliability of a method is its ability to provide the same answer in repeated observations. (The terms reliability and *precision* are often used interchangeably.) Reliability has two aspects: inter-interpreter and intra-interpreter.

For evaluation of inter-interpreter reliability (consistency of observations by different interpreters—in our example, radiologists), a set of scans showing a broad range of abnormalities, and including some showing normality, are presented in random sequence for interpretation by each radiologist participating in the study. Of course the true status of each subject must be unknown to the radiologist at the time he views the scan (but this information should be available for subsequent assessment of the accuracy of scan interpretation). This type of evaluation requires a large number of scans: at least 100, and sometimes more.

It is also desirable to evaluate intra-interpreter reliability (the consistency with which the same interpreter arrives at the same diagnosis when viewing the same scan). This may be accomplished by repetition of the study outlined above. However, any possible learning effect should be minimized. A method often employed to accomplish this, at least in part, is to use a large number of scans, randomly rearrange the order for each repetition, and separate the repetitions by suitably long time intervals. It is

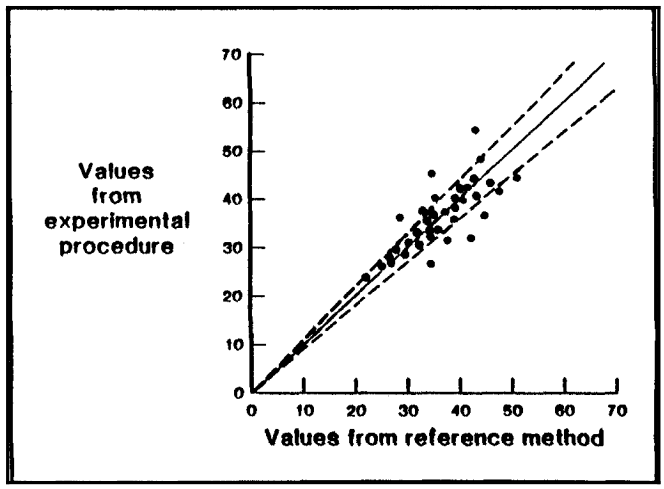

Fig. 9-1. Measurements made by experimental method related to those from reference method, with lines of identity (solid) and ± 10% error (dashed).

important that this study, like the one above, be done without knowledge of the true status of the patient.

Accuracy.The accuracy of a procedure is measured by its ability to give the right answer. Usually this is expressed by the rates of false-positives and false-negatives. Here, of course, it is necessary that the true status of the patients be established. In many studies this will be done by surgery or other definitive diagnostic procedures; in others, the status can be determined by follow-up. Since the willingness of the physician to submit his patient to further study or surgery may be reduced by a negative result of the procedure actually under question (in this example, the computer-assisted scanning), determination of how many negative results are false may become the more difficult problem.

The number of patients required for estimation of the true rates of false-positives and false-negatives depends on how low these rates are, how accurately they need to be estimated, and what proportion of patients in the study have tumors. It is often desirable to use only patients who are scheduled for surgery, thereby including many subjects whose findings are positive and at the same time minimizing the problem of assessing false-negatives, since the surgery will reveal the true status in each case.

Reasonably accurate estimation of these rates usually requires a large number of subjects—a group of 100 might be considered a small but acceptable sample. In order to determine the adequacy of a contemplated sample size, the investigator must indicate approximately the rates anticipated. The statistician can then indicate how accurately those rates can be estimated from a sample of the size contemplated.

Comparative Studies.—The usual objective of a comparative study is to compare the accuracy of the experimental method (scanning) with the accuracy of one or more conventional methods (such as conventional roentgenography). An important first step is to define the patient group to be studied. It is essential in this type of study that eligibility for the study should not depend in any way on the outcome of either the experimental or the conventional method. For this reason, the patient's entry into the study should be determined before he is examined by either method. Once a patient is admitted to the study, examination by each method should be done without knowledge of the results of the competing method. Additional knowledge (certain clinical information, for example) should not be available with either method (unless such information is considered an integral part of that method).

For ascertaining the relative accuracy of the two methods, the true status of the patients must be known. For example, if method A indicates the presence of a tumor when method B does not, resolution of this difference may be obtained from subsequent surgery. In this situation, the willingness to do surgery should be the same when A is positive and B is negative as when A is negative and B is positive. If it is known beforehand that the rate of false-positives for each method is near zero, this difficulty does not arise.

In the absence of a definitive diagnosis, the best that can be done is to measure *agreement* between methods A and B without attempting to measure relative accuracy.

CONTINUOUS DATA

The concepts of reliability, accuracy, and comparative studies described above still apply when the measurement of interest is continuous—as is tumor size. However, some of the statistical methods are slightly different.

For example, the reliability (consistency) of observations may be expressed by the standard deviation among repeated measurements. It sometimes happens that the error tends to be larger when the quantity under study is large—errors may tend to be larger in measuring a large tumor than a very small one, for example. To counter this, it may be appropriate to express reliability by the *coefficient of variation,* which is the standard deviation divided by the mean ($s \div \bar{x}$).

In measuring accuracy, we are concerned with how closely repeated measurements cluster about the true value. Sometimes closeness is best measured by the algebraic difference from the true to the observed value. When the difference seems to be proportonal to the magnitude of the true value, it may be more appropriate to express the difference as a percentage of the true value. Often a graph such as that shown in Figure 9-1 is helpful

in evaluating accuracy. A graph of the cumulative distribution of the error (expressed either as a difference or as percentage error) also may be useful, or perhaps quoting appropriate percentiles from the cumulative distribution will suffice. Sometimes the absolute magnitudes of the error (in which negative signs are disregarded) are most informative.

When results from two methods of measurement are to be compared and the definitive measurement is available (such as tumor size determined at surgery), one can tabulate the errors for each method and compare the two distributions of error. A statistical test of significance also may be performed (based on the values of the observed errors—perhaps using a paired Student *t* test). However, statistical significance ordinarily is of minor concern in evaluating a new procedure.

COMMENT

Once the descriptive techniques described above (and perhaps others) have been employed, the ultimate question, *are the reliability, the accuracy, and the improvement over existing methods good enough?* must be answered by the physicians.

Statistics for Clinicians

10. Normal Values

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

In interpreting the significance of a laboratory measurement, it is often helpful to know how the value obtained in a given case relates to a set of values from a healthy reference population. What percentage of healthy persons have higher (or lower) values?

For this purpose, we must find the distribution of the variable in the healthy population; and that process is the topic of this paper.

ESTIMATING NORMAL VALUES

Distribution in Sample.—The basic approach is the same as in several previous undertakings: since it is not possible to make observations of every member of the population, we rely instead on estimates derived from a sample. For illustration, consider the population of 5,594 serum urea values in paper 4 (Estimation From Samples). Since those values were unknown to us, we drew a sample of 100 values randomly from the population with which to estimate characteristics of the population (such as its mean). The same values are presented again in Table 10-1 with percentile values added.

As usual, the high percentiles are matched to the large values, and the percentile matched to a value indicates what percentage of all the values rank lower. Thus the largest of the 100 values (173 mg/dl) is the 99th percentile (P_{99}); the next value (103 mg/dl) is the 98th percentile; and so on. If there were 200 values in the sample, rather than 100, the largest observation would estimate the 99.5 percentile; and if the 10th largest value were still 82 mg/dl, then that would still be our estimate of the 95th percentile.

As you may have noticed in Table 10-1, 95% of the observations were less than 82 mg/dl; but 95% were also less than 69, or any number between 82 and 69. As a result, any of these numbers could be used to estimate the 95th percentile. Rather than choosing the largest, it is conventional to choose a value in between (such as 75). Various strategies for making an appropriate choice have been developed and are commonly used. In general, larger samples produce smaller gaps; and sample sizes should be made large enough so that the ambiguity resulting from this problem is negligible.

Provided with a distribution of percentile values in a sizable sample, a physician can determine approximately how his patient's serum urea value relates to those in the reference population. From Table 10-1, for example, he would know that a value as large as 50 mg/dl is uncommon—estimated to occur in only 10% of that population.

Reliability (and Sample Size).—As in any situation where we must rely on sample estimates, we are concerned with their reliability. Here we consider percentile estimates from 10 samples of 100 each, drawn from the population of 5,594 (the same samples drawn in paper 4, now represented by selected percentile values in Table 10-2). Clearly, the

Table 10-1.—**Distribution of Serum Urea Values in a Sample (n = 100) Drawn Randomly From a Population (N = 5,594)***

Value, mg/dl	Frequency and percentile (P)	Value, mg/dl	Frequency and percentile (P)
173	1	36	2
103	1	35	6
95	1	34	2
88	1	33	3
82	1 (P_{95})	32	9 (P_{50})
68	1	31	4
66	1	30	6
52	2	29	6
50	1 (P_{90})	28	2
46	1	27	2 (P_{25})
45	2	26	2
44	1	25	4
42	5	24	6
41	3	23	3 (P_{10})
40	5 (P_{75})	22	2
39	2	20	5 (P_5)
38	1	19	1
37	3	18	1
		16	1

*Mean of sample is 36.56; standard deviation is 20.27. (Data same as in Table 4-1, with percentiles added.)

Table 10-2.—**Mean and Selected Percentiles of Serum Urea Values in 10 Samples (Each n = 100)***

Sample	Mean, mg/dl	Values for selected percentiles			
		P_{50}	P_{90}	P_{95}	P_{99}
1†	36.56	32	50	82	173
2	33.92	31	50	57	103
3	34.24	31	50	62	123
4	33.00	31	43	52	86
5	33.47	31	46	60	220
6	36.67	32	48	56	172
7	35.15	30	52	61	123
8	38.93	32	50	69	388
9	32.31	30	48	56	93
10	36.57	32	46	55	174
s‡	2.07	0.8	2.7	8.8	89.3
Population values§	35.33	31	48	60	124

*Samples presented in Table 4-2.
†From Table 10-1.
‡The standard deviation (s) of the 10 values listed directly above.
§From population of 5,594 values (paper 4).

values for P_{50} are more uniform than the values for the very high percentiles (P_{90}, P_{95}, and P_{99}). Although a sample consisting of 100 values ordinarily is adequate for estimating the center of a population, it is a very small basis for estimating the outer percentiles (such as P_5 or P_{95}).

Refinements.—In our example we have deliberately oversimplified the problem of estimating normal percentiles. Normal values of many variables are affected by the age and sex of the subjects. Statistical methods are available for estimating age- and sex-specific percentiles, but they obviously require data from more subjects.

COMMENT

As mentioned in paper 1, it is a common misconception that, in general, 95% of population values lie within two standard deviations of the population mean. (The proposition is true only under special, infrequently occurring conditions.) This misconception has given rise to the regrettable practice of estimating the 2.5 and 97.5 percentiles simply as the mean ± 2 standard deviations ($\bar{x} \pm 2s$). Applied to the first sample of 100 serum urea values in our example (presented in Table 10-1), $x \pm 2s$ yields $36.56 \pm 2 \cdot 20.27$, giving the impossible result $P_{2.5} = -3.98$ mg/dl. Clearly, the method is unsuitable for general use.

For two nontechnical papers providing an excellent, more detailed discussion regarding the choice of a suitable reference population and the estimation of population percentiles, see Elveback.[1,2]

REFERENCES

1. Elveback LR: How high is high? A proposed alternative to the normal range. Mayo Clin Proc 47:93-97, 1972
2. Elveback LR: The population of healthy persons as a source of reference information. Hum Pathol 4:9-16, 1973

Statistics for Clinicians

11. Survivorship Studies

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms introduced
Direct (ad hoc) analysis
Actuarial (life-table) analysis
Observational study
Experimental study

Definition of Study Group.—In every study of survivorship—as in virtually all medical research on human subjects—the first requirement is to describe the group studied. The reader of the report must be told the nature of the group so he can judge whether his patient or group is like it. The description should include:

1. The source of subjects and the period in which they entered the study, with notice of any considerable selection bias (practice in a general hospital or a specialty clinic, and so forth).
2. The medical problem of interest: what it was, and how its presence was determined. In some studies it is desirable to distinguish subtypes of the problem, or degrees of severity.
3. The treatment, if any.
4. All exclusions of subjects from the study, and the reasons for them.
5. Characteristics of the study group: their age and sex distributions; if pertinent, their area of residence, occupations, economic status, and race; and so on.
6. Complicating features (associated diseases, and so forth) if it seems they may affect survival.

DATA COLLECTION AND ACCOUNTING

Completeness of Follow-Up.—The problem in follow-up is the practical difficulty of making it complete enough. Much effort and many stratagems may be justified, because a case "lost to follow-up" cannot be ignored. Even if entirely excluded from the analysis, it must be mentioned in the report and remembered in judgment, because cases lost may not have had the same outcome as the cases traced. No amount of sophisticated mathematical manipulation can overcome failure of follow-up in a sizable number of instances.

Initial Event.—In survivorship studies, each case must have an initial event from whose date the period of observation is counted. This may be birth, for congenital disease; but usually it is diagnosis, surgery, or beginning of treatment. Although the onset of disease might be very meaningful, dating of onsets is often difficult. Surgical and hospital deaths may be excluded (if exclusion is desired) by beginning at a time such as "30 days after operation."

Accounting of Follow-Up Period.—Since the initial event does not occur simultaneously in all cases, the lengths of follow-up are not equal at any given date. Survivorship analysis is based on an equal follow-up

Table 11.1—**Minimal Information to be Recorded for Study of Survivorship**

1. Sex
2. Date of birth (to give age at initial event)
3. Date of initial event
4. Date of latest follow-up (of death, if dead)
5. Status at latest follow-up (dead or alive)
6. Cause of death (if available)

interval, however—which is attained at different times, case by case. A subject becomes eligible for inclusion in analysis of survival for a given period when that much time has passed since the initial event in his case. Thus a patient whose cancer was resected 3 years ago is eligible for inclusion in analysis of 3-year survival, despite having died of recurrence 2 years after the resection. In 2 more years he will become eligible for 5-year analysis; but he is not eligible for it now, even though we know now what his status will be then. To advance a 3-year nonsurvivor to the 5-year calculation would unbalance it, because we do not know how to advance (as alive or dead?) the other 3-year subjects presently surviving—who must be considered with him.

Data Collected.—The minimum of information on each subject for routine statistical analysis is listed in Table 11-1.

ANALYSIS OF DATA

Direct (Ad Hoc) Analysis.—Direct determination of a survival rate is done with this formula:

$$\frac{\text{Subjects who survived through the period of observation}}{\text{Subjects who survived that long plus those who were eligible but died}}$$

Single-Period.—Some years ago, it was usual to analyze survival data for the 5-year rate alone. For example, if gastrectomy had been performed on 84 patients 5 or more years previously, and at 5 years after operation (case by case) there were 42 surviving, the 5-year survivorship was 42/84 = 50%.

However, single-period analysis has two major inadequacies. First, a single-period rate does not reveal survivorship at any time preceding or following the end of the period chosen. Second, a single-period analysis (unless the period is brief) excludes a great deal of data an investigator is likely to have accumulated from more recent cases.

Serial Determinations.—It is possible, of course, to perform direct-method calculations on periods expanding from the initial event (1-year, 2-year, 3-year, for instance; not first-year, second-year, third-year), each time using all the cases eligible for the period being considered then. These serial determinations should reveal any trend within the maximal period analyzed.

However, the resultant series of rates may not be very accurate. Indeed, if there has been less mortality among early cases than recent ones, this method may produce survival rates that rise with the length of follow-up; and some degree of such distortion may be present without being obvious. And each determination still excludes data from the computations. Therefore this method is often not a good choice. For a more detailed nontechnical discussion, see Berkson and Gage.[1]

Actuarial (Life-Table) Analysis.—Typically more accurate than the direct method is the actuarial method. This is based on the question, applied to each day of observation n (n = 1, 2, . . .): "For subjects who survived n days, what is the probability (P_n) of surviving one more day?" (To estimate this probability, we divide the number of subjects who actually survived $n + 1$ days by this number plus the number who died on the $n + 1st$ day.) The probability of surviving from day 1 through day n is then estimated by the product of the probabilities of surviving each day ($P_1 \cdot P_2 \ldots P_n$). Although the computations for this method may appear cumbersome in computing a 5-year survival rate, they are greatly simplified by the fact that, except for days on which deaths occurred, $P_n = 1$. (And typically, for very large data sets, the computations are performed by computer.)

The major advantage of the actuarial method is that it utilizes all the available data: every subject is counted for whatever time he has been followed, no matter how brief. This makes the estimated survival rates more reliable. Second, the rates for successive intervals are combined in a way that excludes distortion. A curve that makes survival appear to increase as time passes is not possible.

Deaths Due to Unrelated Causes.—Thus far we have described determination of the gross death rate among a study group. However, if any of the deaths were due to causes other than the risk factor under study—and if the investigator is sure of his knowledge in every case—he must decide whether to determine and report the cause-specific death rate. This is accomplished by treating as deaths only those instances caused by the risk factor. Unrelated deaths are treated as lost to follow-up at time of death. Usually the particular study dictates the greater interest, and sometimes both rates are of interest.

PRESENTATION OF RESULTS

Generally the most effective method for describing the survival experience of a group of patients is to graph survival rates against time, as shown in Figure 11-1.

To provide perspective on the outcome of an analysis, a comparison with normal survivorship may be shown. The appropriate norm is experience in a segment of the general population, adjusted (from published tables) to match the study group in respect to age, sex, and perhaps other features that seem pertinent. These rates will indicate the survivorship that would have been expected in the study group if it were representative of the general population. Additionally, expected 5-year or 10-year rates might be presented in the text.

COMMENT

The principal concern of this paper is to point out the need to take varying lengths of follow-up into account in studying survivorship. We hope that has received sufficient emphasis above.

Two other ideas remain for presentation here.

1. The methods for analyzing survival data have been developed more recently than the other statistical methods we have presented, and still newer techniques are being proposed continually. Procedures are available for testing the differences between two or more survival curves, for testing the association between survival and a continuous risk factor (such as the serum cholesterol concentration), and for performing such tests after adjustments for other relevant factors.

2. Interpretation of results is often difficult, however. Because survivorship studies generally are *observational* rather than experimental, questions arise regarding what has caused the differences that are found.

To illustrate, suppose that two different surgical techniques were used to treat patients having the same disease and that 10-year follow-up was obtained on all patients treated with each method. It would be tempting to attribute any difference in survivorship to the difference in surgical techniques. Such a conclusion might not be valid, however, since the disparity could be a result of other factors. For example, the two groups of patients may have been dissimilar with respect to factors that influence the choice of surgical technique (possibly severity of the illness or age of the patient). Unfortunately, sophisticated statistical algorithms are of only limited usefulness in attempts to distinguish effects due to the factor of interest (surgery) from effects due to other causes.

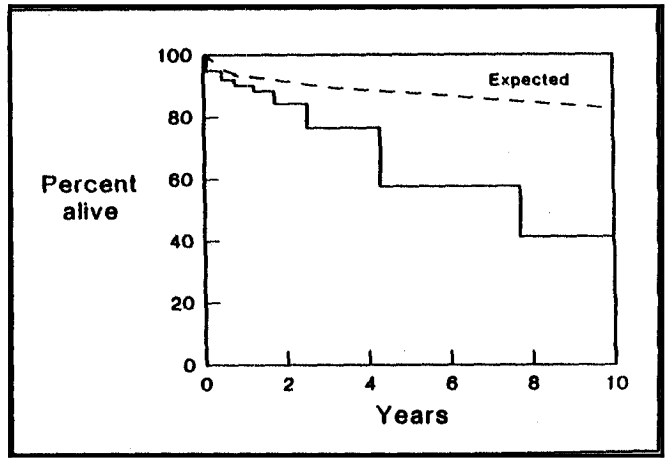

Fig. 11-1. Survivorship (actuarial analysis): as observed in study group (*solid line*) and as derived from population segment similar in regard to age, sex, and date of birth (*broken line*).

In order to establish the relative merits of the two surgical techniques, it would be best to design an experiment specifically with this purpose in mind. Ideally, patients would be assigned randomly to either method, enabling a statistician to make a valid probability statement in comparing the two procedures.

Notice that this was the approach in the *experimental* studies described previously in this series. For example, the experimental study described in paper 6 (Comparing Two Samples) was designed carefully, in advance of data collection, so that a direct comparison could be made of the reduction of blood pressure by each of the two drugs used. When a difference between drug effects is observed in a properly designed experimental study, we can make a valid probability statement regarding the hypothesis that it was caused entirely by other factors instead.

Thus, although an observational study is often considerably more convenient and less expensive than a carefully designed experiment, one must also consider the quality and interpretability of the results ultimately to be obtained.

REFERENCE

1. Berkson J, Gage RP: Calculation of survival rates for cancer. Proc Staff Meet Mayo Clin 25:270-286, 1950

Statistics for Clinicians

12. Sequential Methods

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

Technical terms introduced
Sequential
Fully sequential
Group sequential

In this paper we shall consider whether headache is more frequent with drug F or with drug G—which is the same question that was addressed in paper 8, but our method will be different. For the example developed in paper 8, the study was begun by randomly assigning half of a series of patients to receive drug F and the other half drug G; and when the observations had been made, the investigator tested the hypothesis that the frequency of headache with each drug was the same, using an appropriate statistical method.

This time, however, we wish to monitor the data as they are being collected, with a view toward terminating the trial early if either drug appears definitely superior to the other. As in our previous papers, if the data lead us to conclude a difference—whether at an interim review (with consequent termination of the trial) or at completion—we will want to know the corresponding *P* value. That is, if no real difference existed and our trial were repeated many times, what proportion of those trials would provide such strong evidence of a difference? However, for reasons to be discussed later (in the Comment), the testing methods described in paper 8 are not valid for use with a sequential evaluation of the data: modifications are required.

All sequential methods that have been developed use objective predetermined criteria for termination of the trial. To illustrate, let us suppose that the investigator in our hypothetical example decided he would be willing to study a maximum of 120 patients in a clinical trial, 60 to receive drug F and 60 drug G by random assignment. He plans to evaluate the data when each increment of 40 observations becomes available. At each evaluation, he will use the methods described in paper 8 to compute a chi-square (χ^2) statistic. If any of these statistics is sufficiently large, the trial will be terminated with the conclusion that one drug is superior to the other.

How large is "sufficiently large"? To ensure that the conclusion of a difference will not be reached erroneously in more than 5% of such studies, specially prepared tables (*not* the tables of the χ^2 distribution referred to in paper 8) must be used. When statistical significance is indicated, the tables also provide the corresponding *P* value. These tables indicate that the first 40 cases should yield a χ^2 value exceeding 11.8, or the first 80 cases 5.9, or 120 cases 3.94.

Suppose that in the first group of 40 patients, headache is reported by 5 of the 20 who received F and by 12 of the 20 who received G. These data yield a χ^2 value of 5.0. Since this is less than 11.8, the observed difference

between F and G is not sufficient to warrant stopping the study at this point.

Therefore a second group of 40 patients is enrolled and observed; and the combined results of the two groups are headache in 10 of 40 who received F and in 22 of 40 who received G. These numbers result in a χ^2 value of 7.5; and since 7.5 is greater than 5.9, the evidence at hand is sufficient to warrant termination of the study with small risk that further data would negate the apparent superiority of drug F.

COMMENT

1. In this example, it might have been tempting to compare each observed χ^2 value to percentiles of the tabled χ^2 distribution, as in paper 8. With this strategy, one would have obtained a P value of 0.025 at the first test and—since this is less than 0.05—would have concluded that the difference was statistically significant. But how often will an experimenter using this strategy reject the null hypothesis incorrectly?

By definition, the probability of obtaining a statistically significant result ($P<0.05$) at the first review is 0.05. However, the probability of obtaining this result on review of groups 1 and 2 combined (but not group 1 alone) is 0.033; and the probability of obtaining it on review of groups 1, 2, and 3 combined (but not group 1 or groups 1 and 2 combined) is 0.024. Since the null hypothesis will be rejected under any of these three circumstances, the probability of rejection is 0.050 + 0.033 + 0.024, which equals 0.107.

The reader should remember that, if one makes sequential evaluations of data, special methodology should be supplied by a statistician.

2. The term *sequential,* in statistics, refers to the approach to study design and data analysis in which the data are reviewed at various points during the course of the study. Procedures have been developed for performing a test of significance as each observation is added to the accumulated evidence, but they are generally impractical and rarely used. Such plans are often referred to as *fully sequential.* On the other hand, the type of sequential design that we have described (where a test is performed as successive groups of observations are added to the accumulation) is referred to as *group sequential.*

Epilogue

PETER C. O'BRIEN, Ph.D.
Department of Medical Statistics and Epidemiology

MARC A. SHAMPO, Ph.D.
Section of Publications, Division of Education

In the preceding series of papers, we have described some of the most elementary concepts and methods in statistics. We started with *descriptive statistics,* discussing methods for describing a data set by use of such descriptors as the mean and standard deviation, median, and range (and interquartile range). Graphic techniques for providing a quick visual impression of the data, such as histograms and scatter diagrams, were presented also.

We then turned our attention to *inferential statistics*—establishing generalizations about a population by use of a sample drawn from it. This process was illustrated in a series of papers describing some of the more common techniques, such as confidence intervals, *t* tests, and chi-square (χ^2) tests. In each situation, the basic approach is the same: first, the questions being addressed must be identified and stated precisely. These questions, together with the resources available to the investigator, determine the appropriate study design, which in turn dictates the method used for data analysis. Proper interpretation of the analysis completes the process. It is essential that an investigator who intends to rely on statistical inference work closely with a statistician during the entire process: from questions to study design to data analysis to interpretation.

Two complementary aspects of data analysis were presented: estimation and hypothesis-testing. *Estimation* is attempting (by use of sample data) to ascertain some characteristic of the population, such as the mean serum urea level, or the difference between sets of paired data (such as blood-pressure measurements made before and after treatment, case by case), or the difference between the incidence of side effects associated with two drugs. Because the estimates are based on sample data—which are subject to random variation—we have shown how to assess their precision by deriving standard errors and confidence limits. Since precision improves with increase of sample size, a confidence interval may be viewed as a measure of the adequacy of sample size.

For *hypothesis-testing,* one first transforms the question of interest into a null hypothesis. For example, to determine whether a new treatment modality is more effective than the established modality, one formulates a hypothesis that there is no difference between their effects. To assess the null hypothesis, one collects data and computes a *P* value. Rejection of the null hypothesis is based on a statement such as, "If the null hypothesis (no difference) is true of the population, then the probability that a sample of this size will show a difference as large as the one that appears in our sample is less than *P*."

When the data justify rejection of the null hypothesis (that is, when the *P* value is very small), the results are termed *statistically significant* (not to be confused with *clinically significant,* a judgment to be made by a clinician). When the results of hypothesis-testing do not lead to rejection of the null hypothesis, the interpretation may be less clear. Accepting the null hy-

pothesis may not be justified if the lack of statistical significance may be attributed to small sample size. Again, this question may be addressed by consideration of confidence intervals, if available.

An important principle is that statistics can only establish an association and cannot define the cause and effect. For example, statistics may establish an association between having a yellow-stained index finger and the occurrence of lung cancer. However, it is obvious that although the association is strong, "yellow finger" does not cause cancer. In this case the observed association between yellow finger and lung cancer is merely an artifact resulting from the association between smoking and lung cancer.

Some additional special topics that occur commonly in medical research were discussed: evaluating a new diagnostic procedure, determining normal values, describing survivorship, and finally, using sequential methods. Although we alluded only briefly to some important study-design considerations, it is worthwhile to keep in mind the need for a comparison group, the desirability of random double-blind treatment assignment, and the important distinction between observational and experimental studies.

In all the topics introduced, we only scratched the surface; and of necessity, some topics were omitted entirely. However, we hope we have provided the reader with an introduction that will encourage a further study of statistics and prepare him for wiser judgment of what he reads.